Do Hesitations Facilitate Processing of Partially Defective System Utterances? An Exploratory Eye Tracking Study

Kristin Haake[1], *Sarah Schimke*[1], *Simon Betz*[2], *Sina Zarrieß*[2]

[1]TU Dortmund University
[2]Universität Bielefeld

kristin.haake@tu-dortmund.de, sina.zarriess@uni-bielefeld.de

Abstract

Spoken dialogue systems are predominantly evaluated using offline methods such as user ratings or task-oriented measures. Various phenomena in conversational speech, however, are known to affect the way the listener's comprehension *unfolds* over time, and not necessarily the final result of the comprehension process. For instance, in human reference comprehension, conversational signals like hesitations have been shown to ease processing of expressions referring to difficult-to-describe targets, as can primarily be observed in listeners' anticipatory eye movements rather than in their final reference resolution decision. In this study, we explore eye tracking for testing conversational dialogue systems, looking at how listeners process automatically generated referring expressions containing defective attributes. We investigate whether hesitations facilitate the processing of partially defective system utterances and track the user's eye movements when listening to expressions with: (i) semantically defective but fluently synthesized adjectives, (ii) defective and lengthened adjectives, i.e. containing a conversational uncertainty signal. Our results are encouraging: whereas the offline measure of task success does not show any differences between the two conditions, the listeners' eye movements suggest that processing of partially defective utterances might be facilitated by conversational hesitations.

Index Terms: spoken dialogue, conversational systems, reference comprehension, synthetic speech processing, hesitations, speech synthesis, eye tracking

1. Introduction

In research on spoken dialogue systems, it has long been recognized that even highly fluent speech synthesis and generation is not enough to always achieve naturally flowing interactions with human users. Beyond fluency, dialogue systems should be equipped with capabilities of *conversational* speech production in order to account for various phenomena of spontaneous human dialogue such as corrections, pauses, interruptions or hesitations [1, 2, 3]. These conversational capabilities would be particularly useful when the dialogue system is, for some reason, not able to perfectly deliver a specific utterance, e.g. when it has not finished processing, when it is uncertain, encounters noise, etc. [4]. Even though past years have seen increasing interest in conversational dialogue modeling, it is notoriously hard to test these systems and assess their strengths in comparison to traditional systems that typically deliver fluently synthesized, but schematic and less adaptive speech output [5]. For evaluating speech synthesis and spoken dialogue systems, offline methods such as user ratings or task-oriented measures remain predominant. Various phenomena in conversational speech, however, are known to affect the way the listener's comprehension *unfolds* over time, rather than the final result of the comprehension process. Thus, when relying on offline methods for testing conversational systems, it is possible that the effects of relatively subtle aspects of conversational speech (e.g. pauses or hesitations) might not fully be reflected in the evaluation results.

Figure 1: *Example of a visual stimulus and a partially defective generated referring expression for the target object*

In this paper, we present an exploratory study that uses eye tracking to investigate whether a human listener has advantages in processing conversationally synthesized utterances generated by a dialogue system, as compared to schematic synthesis. The dialogue system's task is to generate expressions referring to objects in real-world images, intending to identify these objects to a human listener. The system's underlying generation component predicts words directly from low-level visual input representations of the target object defined via a bounding box in the image, based on the approach in [6, 7]. As illustrated in Figure 1, this can lead to partially defective utterances being generated, due to imperfect visual language grounding [7]. We investigate whether hesitations facilitate the processing of partially defective system utterances: psycholinguistic studies on human reference resolution show that listeners react very quickly to paralinguistic markers such as hesitations during comprehension and that these reactions may be reflected in listeners' eye movements ([8, 9, 10]).

In the current study, we investigate whether this is the case when hesitations are produced by a computer system in cases of uncertainty about the correct color term for a given object. Thus, we track the eye movements of users listening to expressions with (i) semantically defective but fluently synthesized adjectives, (ii) defective and lengthened adjectives, i.e. containing a conversational uncertainty signal. Our results suggest that indeed, listeners may have an online advantage at identifying the intended object in cases where defective color terms were produced with hesitations. Importantly, we did not observe an effect of hesitations in the offline accuracy in identifying the intended object, suggesting that eye tracking may be a more sensitive measure and a useful complement to offline measures when evaluating dialogue systems.

2. Background and Overview

Dialogue systems for real-world reference games Reference games are a widespread experimental setup for psycholinguis-

tic studies and a popular domain for generation or dialogue systems that target interactions with users about a visual scene, cf. [11]. While traditional work on reference generation looked at relatively simple scenes with graphical objects and focused on restricted tasks like attribute selection, research on language generation has recently started to investigate setups based on real-world images [12, 13, 7]. Here, the system's task is to generate a semantically and pragmatically adequate expression given only the low-level visual features of an image region. This setting poses new challenges for dialogue systems as, for instance, these systems might produce relatively disturbing errors like defective nouns (e.g. the system generates *tree* vs. *man*) due to perceptual uncertainty [7]. Thus, it has been argued, that dialogue systems interacting with users in real-world environments need principled communicative mechanisms for dealing with uncertainties, perceptual mismatches, and potential misunderstanding [14, 15, 16, 17]. In this study, we look at (potentially less disturbing) defective color terms, which turn out to be hard to predict for objects in real-world images as well [6]. Thus, for compiling the materials of our experiment, we used color terms predicted for objects in images by [6]'s model which we identified as defective based on annotated color terms in the training set of the model. This allows us to investigate whether conversational synthesis can lead to processing advantages for the listener for realistic, but controlled errors in system output.

Hesitations in conversational speech synthesis Generally, hesitations in natural conversational speech can be realized in terms of different phenomena allowing the speaker to temporally extend the delivery of the message, i.e. lengthening of phones and syllables, silent pauses or production of content-free material, such as fillers (*uh, uhm*) or non-committing words (*like, you know*) [18]. While hesitations in human interaction are relatively well-studied (see Section **??**), it is less clear whether and how they should be modeled in human-machine-communication. On the one hand, it has been shown that users of conversational dialogue systems can perceive synthetic hesitation, mainly caused by deviations from expected temporal patterns [19]. On the other hand, findings concerning the effect of hesitations on interaction quality are mixed: [20] found that synthesized fillers have no positive effect on comprehension; [5, 21] found that synthetic lengthening does enhance users' task performance and efficiency. Furthermore, the relation between hesitation and its phonetic realization is subject of ongoing research. [22] found that users prefer a system that deploys fillers, while [20, 23] explicitly point out difficulties of filler synthesis compared to other hesitation types. Lengthening and silence fare well in terms of user feedback in offline evaluation studies [23], but embedded in a smart-home scenario, systems that hesitate only by means of silence are perceived as less friendly than non-hesitating systems [24, 25]. For the purposes of this study, we focus on lengthening as a marker of hesitation and uncertainty. Lengthenings are promising for our purposes because they can be realized directly on the defective attribute and do not seem to significantly degrade the perceived quality of the synthesis [23].

Processing of hesitations in natural speech Psycholinguistic studies have shown that listeners are able to quickly perceive and integrate paralinguistic signals, such as hesitations, into their current understanding of an unfolding utterance. [8] used the visual world paradigm to study how listeners process fluent and disfluent descriptions of discourse-new (not previously mentioned) and discourse-old (previously mentioned) objects (e.g. camel and candle). In this paradigm, participants are presented with visual displays containing several objects, and their eye movements to these objects are monitored while speech is unfolding. This allows for assessing subtle effects of linguistic and paralinguistic cues during the processing of speech (see [26]). In [8]s' study, participants listened to either fluent descriptions (e.g. *put the candle*) or to disfluent descriptions (*put thee uh candle*). Their eye movements revealed that disfluencies created a bias towards discourse-new objects before and immediately following the onset of the referring expression. This suggests that disfluencies may facilitate processing by preparing listeners to expect referents that they would not expect otherwise. Hence, the processing of the corresponding referring expressions was facilitated and occurred faster. Later work showed that this also holds for objects that are hard or easy to describe [10], with hesitations apparently signaling to listeners that a hard-to-describe object is the likely next referent.

This study In the current study, we will build on these findings by asking whether hesitations influence online processing also when human listeners have to process unexpected color terms generated by a dialogue system. As alluded to above, we will focus on cases where the system was uncertain about an appropriate color term for an object. We will compare **three conditions** where the systems generates: 1) the matching color term (baseline condition), 2) a defective color term and produces this term without any hesitation, 3) a defective color term, and introduces a hesitation into the production of this term. Importantly, this set-up differs from previous psycholinguistic studies in several ways: We study synthetic speech instead of natural speech, we study real-world images instead of schematized ones, and we study a strong type of mismatch, namely, a defective color adjective. We suspect that all of these differences may lead to comparatively slower reactions in the human listeners when compared to previous studies (see for instance [27, 28]). We ask whether, on top of such a potential general slow-down, there will be differences between defective color terms that are produced with compared to without hesitations. Based on [8], we expect that hesitations may prepare listeners for an unexpected expression, and speculate that this may facilitate recovery from defective color adjectives and speed up the identification of the correct object based on the noun.

3. Experiment

We conducted a visual world paradigm study to test how users process partially defective referring expressions generated by a dialogue system. We examine whether synthetic lengthening of defective color terms facilitates the (presumably necessary) online revision during processing, since the lengthening may be taken as a signal of uncertainty [29].To investigate this, we will analyze the percentage of target looks in the 2000 ms following noun onset, to capture differences in the online revision process from the earliest moment it can possibly be initiated (noun onset) till the noun has presumably been fully processed.

3.1. Design and materials

We created 30 experiment and 40 filler trials, two of which were used as warm-up trials. Each trial consisted of a visual as well as an auditory stimulus. We used real-world images from the IAPR TC-12 corpus [30] for the visual stimuli. Each image depicts a natural visual scene, with a potential target marked by

a red frame (see Fig. 1). Each visual stimulus is composed of 3 images, with a target object, a competitor object, and a distractor object. Importantly, we chose those objects as targets for which [6]'s generation model has difficulties in identifying the correct color, and where the model's best candidate for the color adjective constitutes a defective description. The competitor and distractor objects were selected from the same corpus. For competitors, we used objects (i) of a different type than the target (e.g. "table" vs. pullover", see (1)) and (ii) whose color corresponded to the defective color chosen by the system for the target object (e.g. "yellow" instead of "red" in Fig. 1). This way, the competitor should be the most likely candidate referent for participants listening to the color adjective in the conditions where this adjective was defective. The distractor was chosen to contain an object with neither of these two colors. Because articles and adjectives are gender-marked in German, we made sure that the target and the competitor, but not the distractor, matched in gender, such that gender information would not help distinguishing target and competitor. The objects were presented at different positions on the screen that varied for each experimental trial in a pseudo-randomized fashion. We presented automatically generated and synthesized (using the male voice of MaryTTS ([31]) descriptions of the target object as auditory stimuli. The descriptions were phrased as a search instruction, see (1).

(1) Suche den roten Pullover.
Search the red*(COLOR-ADJ)* pullover*(*NOUN).

We based the phrasing of the stimuli on the object descriptions attested in [32]'s corpus and translated these to German. In condition 1 (baseline), we used the original color term and noun attested in the corpus. In condition 2 and 3, we used the color term predicted by [6]'s generation model.

We conducted a pre-test via crowdsourcing[1] to ascertain that users would prefer the correct color term (condition 1) over the defective term (condition 2 and 3). In this test, we presented images of all target objects with either the correct or the defective color adjective, and elicited image-word compatibility ratings on a six-point Likert Scale from 60 native speakers of German, where "1" meant "strongly agree" and "6" "strongly disagree". The results show that the correct color terms were always rated as more fitting for describing the color of the target (mean: 1.7, variance: 0.9) than the defective color term (mean: 5.52, variance: 0.63).

We created three experimental lists: each participant saw each item in one of the three conditions only, but each item occurred in each condition across the experiment, and each participant was presented with exactly 10 items per condition. We compare three conditions: 1) the correct adjective with regular synthesis 2) a defective color adjective with regular synthesis, and 3) a defective color adjective where the first syllable is lengthened based on inherent phone elasticity [33]. The 40 filler trials were inserted such that participants would not notice a pattern during the experimental trials. In the filler trials, there were no defective descriptions but objects had to be identified based on correct descriptions of their color, size or texture.

3.2. Procedure and Participants

The experiment was programmed using SMI Experiment Center on a DELL Laptop where the auditory stimuli were played as well. The visual stimuli were presented on an external monitor. The experiment started with a synthesized instruction and two warm-up trials that were the same for all participants. The instruction explained the task and provided an option to familiarize with the synthetic voice. Next, the experimental and filler trials were presented in random order to each participant. Each trial started with a preview of 3000 ms before the auditory stimulus was played to make sure that the participants had enough time to perceive the objects in the complex images. After the auditory stimuli, the participants had to click on the chosen object to initiate the next trial. We measured both eye movement during the task and the choice of the participants. Participants were 19 native speakers of German (11 female, M_{age} = 26 years) from the University of Münster.

[1] crowdflower.com

4. Results

The experimental set-up explained in Section 3 allows us to analyze the effect of hesitations in terms of (i) off-line task success that users achieve in the reference game with the dialogue systems, and (i) online processing difficulty as revealed by looks to target and competitor objects. We will discuss these two types of measures in the following.

4.1. Offline measures

To test whether there were significant differences in the number of target clicks – participants' choice of the best matching object – between conditions we run chi2 tests comparing the amount of target (aggregated per participant) clicks in condition 1 compared to 2 and condition 2 compared to 3. There was a significant difference ($\chi^2(1)$ = 13.3, p = .00027) between condition 1 and 2, but not ($\chi^2(1)$ = .8879, p = .346) between condition 2 and 3.

Discussion When selecting a referent, participants mostly relied on the noun describing the object. But they followed the adjective information in about 15 percent of the cases in which they contradicted the noun (in condition 2 and 3). This was not significantly influenced by the synthesis manipulation. This means that while the defective color terms sometimes led to wrong identifications compared to the baseline condition, the more subtle difference between adjectives synthesized with and without hesitations was not reflected in the offline measure. Thus, in a next step, we ask whether there are differences between condition 2 and 3 during online processing. In the examination of the online data we will look at all cases together (e.g. independent of whether ultimately, participants relied on the noun for identifying the correct object) as well as those trials where the target was correctly identified. Thereby we can be sure that the misleading color information from defective adjectives had either been ignored or recovered during online processing.

4.2. Online measures

Figure 2 displays the percentage of looks to the target object out of all three potential objects. Figure 3 displays the percentage of looks to the competitor object. In both figures, the first panel displays the looks for all trials and the second for those where the participants ultimately chose the target, such that we can assume that a revision had to take place following the defective adjective. The x-axis displays the time relative to the onset of the noun for each individual trial, thus, the time point zero represents the onset of the noun for all trials. The figures start with the 500 ms preceding the noun onset. This covers a large part or the entire duration of the adjective. As is clearly visible in

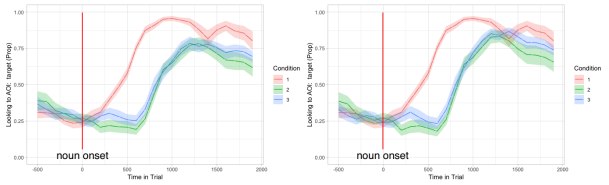

Figure 2: *Proportion of looks on target for all items vs. only items with clicks on target*

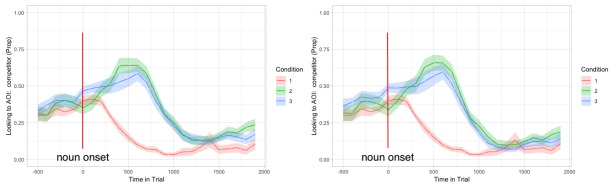

Figure 3: *Proportion of looks on competitor for all items vs. only items with clicks on target*

the graphs, there are no differences between conditions before the noun onset, during the time where participants listened to the adjective, suggesting that any processing of the information contained in the adjective is reflected in the eye movements after the onset of the noun. The time slot from zero to 2000 ms represents the entire duration of the noun and a part of the remaining time of the trial – note that trials ended when participants took a decision (fastest after 2060 ms, in average after 4826 ms).

The results after noun onset show that there is a sharp increase in looks to the target, as well as a decrease in looks to the competitor following the noun onset in condition 1. This suggests that even though during the processing of the adjective, participants did not yet show any behavioral reaction to the information provided by it, they must have processed this information and used it to anticipate the noun, explaining the steep increase in condition 1 compared to the other two conditions, where no such preparation was possible. The data for conditions 2 and 3 reveal that in most trials, participants recovered from the misleading adjective information after noun onset, but that it took them about 700 ms before looks to the target started to increase and looks to the competitor started to decrease in these two conditions.

When comparing condition 2 and 3, interestingly, the revision does not seem to proceed in an exactly identical manner. In the 400 ms following the noun onset, there are more looks to the target after a lengthened adjective than after a regular adjective, and fewer looks to the competitor. This effect, albeit small and transitory, suggests that as expected, a revision may have been facilitated by the preceding hesitation.

To test for significant differences in the relative amount of looks between the three conditions, we used linear mixed effects model using EyetrackingR ([34]). The dependent variables were first, the e-transformed log odds of looking to the target object relative to the other objects and, in a second analysis, the e-transformed log odds of looking to the competitor object relative to the other objects. We added condition as a predictor to the model as well as random intercepts for participants and items and the by-item and by-participant random slope of condition. The analysis was conducted for the entire two second window for the noun onset and revealed a significant difference (all clicks and target looks: estimate = -1.86, SE = .24, t = -7.89, $p < .001$; all clicks and competitor looks: estimate = 2.25, SE = .28, t = 8.003, $p < .001$; target clicks and target looks: estimate = -1.67, SE = .24, t = -7.1, $p < .001$; target clicks and competitor looks: estimate = 2.11, SE = .29, t = 7.358, $p < .001$) between conditions 1 and 2 and a small but significant difference (all clicks and target looks: estimate = .38, SE = .18, t = 2.06, p = .0497; all clicks and competitor looks: estimate = -.47, SE = .23, t = -2.08, p = .0442; target clicks and target looks: estimate = .38, SE = .18, t = 2.08, p = .0432; target clicks and competitor looks: estimate = -.51, SE = .24, t = -2.150, p = .0377) between conditions 2 and 3.

5. General discussion

Our results provide new insights into the processing of synthesized speech in a spoken dialogue system context. Offline, participants' identification of described objects was hindered by a defective adjective, as reflected in the about 15 percent of the cases where ultimately, a non-intended object was chosen. However, there was no difference in offline success between cases where the defective adjective was lengthened compared to fluent. To find out whether more subtle differences can be detected during online processing, we conducted an online eye tracking analysis. These analyses revealed, first, that processing was delayed in comparison to previous studies with natural speech (e.g. [8, 9]). This may, on the one hand, be due to greater difficulties in processing synthesized speech per se, on the other hand, it could be due to the more complex images that were used in the current study when compared to usually schematized stimuli. The delay was reflected in the missing difference between conditions in fixations before noun onset. Immediately after the noun onset, however, there were differences between conditions. In particular, in the baseline condition, the identification of the noun was clearly facilitated by the preceding matching adjective information. Presumably, the adjective information was used to anticipate the noun, explaining the steep increase in looks to the correct object immediately after noun onset. We can thus assume that participants also used the adjective for anticipation in conditions 2 and 3. This must have led to wrongly anticipating the competitor rather than the target object. Eye movements after noun onset thus must reflect a revision process in these two conditions. Our main research question is whether this revision is influenced by a lengthening of the preceding adjective. Our results suggest that this is the case, as overall, there were relatively more looks to the target object and less to the competitor object following a lengthened than a fluent adjective. Importantly, this difference is not only due to trials where participants chose a non-intended object but persisted in trials were the correct object was chosen. This suggests that the lengthening of the adjective led to participants being less distracted by a competing object that matched in color, but not in the type of object, and thus faster and more successful in the identification of the target object. A closer look at the graphs suggests that this effect is rather small and transitory, and mainly occurs in the first 400 ms following noun onset, as well as shortly before participants took a decision. It thus seems important to replicate this effect in further studies, and potentially to look at whether it proves to be a reliable effect across different circumstances. In particular, one may wonder whether the same effect might occur when participants interact with a system using natural as opposed to synthetic speech.

6. References

[1] S. Kousidis, C. Kennington, T. Baumann, H. Buschmeier, S. Kopp, and D. Schlangen, "Situationally Aware In-Car Information Presentation Using Incremental Speech Generation: Safer, and More Effective," in *Proceedings of the EACL 2014 Workshop on Dialogue in Motion*, 2014.

[2] H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen, in *Proceedings of the SIGdial 2012 Conference*, Seoul, South Korea.

[3] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, 2015.

[4] G. Skantze and A. Hjalmarsson.

[5] S. Betz, B. Carlmeyer, P. Wagner, and B. Wrede, "Interactive hesitation synthesis: modelling and evaluation," *Multimodal Technologies and Interaction*, vol. 2, no. 1, p. 9, 2018.

[6] S. Zarrieß and D. Schlangen, "Towards Generating Colour Terms for Referents in Photographs: Prefer the Expected or the Unexpected?" in *Proceedings of the 9th International Natural Language Generation conference*. Association for Computational Linguistics, 2016, pp. 246–255.

[7] S. Zarriess and D. Schlangen, "Easy things first: Installments improve referring expression generation for objects in photographs," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 610–620.

[8] J. E. Arnold, M. Fagnano, and M. K. Tanenhaus, "Disfluencies signal theee, um, new information," *Journal of psycholinguistic research*, vol. 32, no. 1, pp. 25–36, 2003.

[9] J. E. Arnold, M. K. Tanenhaus, R. J. Altmann, and M. Fagnano, "The old and thee, uh, new: Disfluency and reference resolution," *Psychological science*, vol. 15, no. 9, pp. 578–582, 2004.

[10] J. E. Arnold, C. L. H. Kam, and M. K. Tanenhaus, "If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 33, no. 5, p. 914, 2007.

[11] E. Krahmer and K. Van Deemter, "Computational generation of referring expressions: A survey," *Computational Linguistics*, vol. 38, no. 1, pp. 173–218, 2012.

[12] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg, "ReferItGame: Referring to Objects in Photographs of Natural Scenes," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, 2014, pp. 787–798.

[13] D. Gkatzia, V. Rieser, P. Bartie, and W. Mackaness, "From the virtual to the realworld: Referring to objects in real-world spatial scenes," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1936–1942.

[14] R. Fang, M. Doering, and J. Y. Chai, "Collaborative models for referring expression generation in situated dialogue," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[15] J. Y. Chai, R. Fang, C. Liu, and L. She, "Collaborative language grounding toward situated human-robot dialogue," *AI Magazine*, vol. 37, no. 4, pp. 32–45, 2016.

[16] O. Lemon, S. Janarthanam, and V. Rieser, *Reinforcement learning approaches to natural language generation in interactive systems*. Cambridge University Press, 2014, p. 151179.

[17] D. DeVault and M. Stone, *Pursuing and demonstrating understanding in dialogue*. Cambridge University Press, 2014, p. 3462.

[18] S. E. Brennan and M. F. Schober, "How listeners compensate for disfluencies in spontaneous speech," *Journal of Memory and Language*, vol. 44, no. 2, pp. 274–296, 2001.

[19] R. Carlson, K. Gustafson, and E. Strangert, "Cues for hesitation in speech synthesis," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 1300–1303.

[20] R. Dall, M. Wester, and M. Corley, "The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech," 2014, pp. 56–60.

[21] S. Betz, S. Zarrie, and P. Wagner, "Synthesized lengthening of function words - The fuzzy boundary between fluency and disfluency," in *Proceedings of the International Conference Fluency and Disfluency*, L. Degand, Ed., 2017, pp. 15–19.

[22] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "Modelling filled pauses prosody to synthesise disfluent speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4810–4813.

[23] S. Betz, P. Wagner, and D. Schlangen, "Micro-structure of disfluencies: Basics for conversational speech synthesis," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015, Dresden)*, 2015, pp. 2222–2226.

[24] B. Carlmeyer, D. Schlangen, and B. Wrede, "Exploring self-interruptions as a strategy for regaining the attention of distracted users," in *Proceedings of the 1st Workshop on Embodied Interaction with Smart Environments - EISE '16*. Association for Computing Machinery (ACM), 2016.

[25] M. Chromik, B. Carlmeyer, and B. Wrede, "Ready for the Next Step?: Investigating the Effect of Incremental Information Presentation in an Object Fetching Task," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*. Association for Computing Machinery (ACM), 2017, pp. 95–96.

[26] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy, "Integration of visual and linguistic information in spoken language comprehension," *Science*, vol. 268, no. 5217, pp. 1632–1634, 1995.

[27] M. D. Swift, E. Campana, J. F. Allen, and M. K. Tanenhaus, "Monitoring eye movements as an evaluation of synthesized speech," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.* IEEE, 2002, pp. 19–22.

[28] M. White, R. Rajkumar, K. Ito, and S. R. Speer, "Eye tracking for the online evaluation of prosody in speech synthesis: Not so fast!" in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[29] É. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies." in *INTERSPEECH*, 2017, pp. 804–808.

[30] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *International workshop ontoImage*, vol. 2, 2006.

[31] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[32] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798.

[33] S. Betz, J. Voße, and P. Wagner, "Phone elasticity in disfluent contexts," 2017.

[34] J. Dink and B. Ferguson, "eyetrackingr: An r library for eyetracking data analysis," *Available at www. eyetracking-r. com. Accessed July*, vol. 6, p. 2017, 2015.